

5     **EFFICIENT TESTS OF ASSOCIATION FOR QUANTITATIVE TRAITS AND AFFECTED-UNAFFECTED**  
**STUDIES USING POOLED DNA**

**Related Application**

      This application claims priority to U. S. Ser. No. 60/238,381, filed October 6, 2000  
[21402-139] which is incorporated herein by reference in its entirety.

10    **Background of the Invention**

      The complex diseases that present the greatest challenge to modern medicine,  
including cancer, cardiovascular disease, and metabolic disorders, arise through the interplay  
of numerous genetic and environmental factors. One of the primary goals of the human  
genome project is to assist in the risk-assessment, prevention, detection, and treatment of these  
15    complex disorders by identifying the genetic components. Disentangling the genetic and  
environmental factors requires carefully designed studies. One approach is to study highly  
homogenous populations (Nillson and Rose 1999; Rabinow, 1999; Frank 2000). A recognized  
drawback of this approach, however, is that disease-associated markers or causative alleles  
found in an isolated population might not be relevant for a larger population. An attractive  
20    alternative is to use well-matched affected-unaffected studies of a more diverse population

      Even with a well-matched sample set, the genetic factors contributing to an aberrant  
phenotype may be difficult to determine. Traditional linkage analysis methods identify  
physical regions of DNA whose inheritance pattern correlates with the inheritance of a  
25    particular trait (Liu 1997; Sham 1997, Ott 1999). These regions may contain millions of  
nucleotides and tens to hundreds of genes, and identifying the causative mutation or a tightly  
linked marker is still a challenge. A more recent approach is to use a sufficiently dense  
marker set to identify causative changes directly. Single nucleotide polymorphisms, or SNPs,  
can provide such a marker set (Cargill et al. 1999). These are typically bi-allelic markers with  
30    linkage disequilibrium extending an estimated 10,000 to 100,000 nucleotides in heterogeneous

human populations (Kruglyak 1999; Collins et al. 2000). Tens to hundreds of thousands of these closely spaced markers are required for a complete scan of the 3 billion nucleotides in the human genome. Because each SNP constitutes a separate test, the significance threshold must be adjusted for multiple hypotheses ( $p\text{-value} \sim 10^{-8}$ ) to identify statistically meaningful associations. Consequently, hundreds to thousands of individuals are required for association studies (Risch and Merikangas 1996).

The most powerful tests of association require that each individual be genotyped for every marker (Fulker et al. 1995, Kruglyak and Lander 1995, Abecasis et al. 2000, Cardon 2000) and remain far too costly for all but testing candidate genes. An alternative that circumvents the need for individual genotypes, related to previous DNA pooling methods for determination of linkage between a molecular marker and a quantitative trait locus (Darvasi and Soller 1994), is to determine allele frequencies for sub-populations pooled on the basis of a qualitative phenotype. Populations of unrelated individuals, separated into affected and unaffected pools, have greater power than related populations. Limited guidance has been provided, however, regarding the sample size requirement of tests using pooled DNA relative to individual genotyping, or the efficiency of tests based on a quantitative phenotype relative to an affected/unaffected design.

The phenotypes relevant for complex disease are often quantitative, however, and converting a quantitative score to a qualitative classification represents a loss of information that can reduce the power of an association study. The location of the dividing line for affected versus unaffected classification, for example, can affect the power to detect association. Furthermore, pooling designs based on a comparison of numerical scores are not even possible with a qualitative classification scheme. These distinctions can be especially relevant when populations contain related individuals and qualitative tests have a disadvantage (Risch and Teng 1998).

When performing risk assessment to determine whether a person suffers from or is at risk of developing a complex disorder often requires measuring an underlying quantitative phenotype. Association studies in unrelated populations can implicate genetic factors contributing to disease risk, and experiments using pooled DNA provide a less costly but necessarily less powerful alternative to methods based on individual genotyping. Association

10500T 644E 2650

studies require markers in linkage disequilibrium with causative genetic polymorphisms. Although the sample sizes required for pooling and individual genotyping studies have been compared in certain instances, general results have not been reported in the context of association studies, nor have there been clear comparisons of pooling based on quantitative and qualitative (affected/unaffected) phenotypes. Association tests of DNA pooled on the basis of a quantitative phenotype are analogous to selection experiments for quantitative trait locus (QTL) mapping. For a QTL with a weak effect on a phenotype, the mean phenotypic value of individuals selected to exceed a threshold is proportional to the mean allele enrichment. This suggests that genotyping of a certain percentage of the upper and lower phenotypic values of an unrelated population is useful to estimate the effect of a marker on a quantitative phenotype, such as in pooling studies. There is a need in the art to examine the sample size requirements of association tests for quantitative traits using pooled DNA.

#### Summary of the Invention

The present invention is based, in part, on the discovery of methods to detect an association in a population of individuals between a genetic locus and a quantitative phenotype, where two or more alleles occur at a given genetic locus, and the phenotype is expressed using a numerical phenotypic value whose range falls within a first numerical limit and a second numerical limit. These limits are used to provide for subpopulations that consist of upper and lower pools.

In some embodiments, the population of individuals includes individuals who may be classified into classes. In certain aspects of the invention, these classes are based on age, gender, race, or ethnic origin. In other aspects, some or all members of a class are included in the pools.

In various embodiments, these numerical limits are chosen so that the upper pool includes the highest 19%, 27%, or 37% of the population. In other embodiments, the numerical limits are chosen such that the lower pool includes the lowest 19%, 27%, or 37% of the population.

In some embodiments, the upper and lower pools have the same number of individuals.

In one embodiment of the invention, the numerical limits are chosen to correlate with error of measurement determinations. In some embodiments, the numerical limit on the error of measurement is about 0.04 or about 0.01.

In some embodiments, methods to detect an association in a population of individuals between a genetic locus and a quantitative phenotype are useful to determine the genetic basis of disease predisposition.

5 In other embodiments, the genetic locus analyzed contains a single nucleotide polymorphism.

In the present invention, the population of individuals can include unrelated individuals.

10 Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In the case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, 15 and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description and claims.

### Brief Description of the Figures

20 Fig. 1. The sample size required to achieve a type I error rate of  $5 \times 10^{-8}$  and a power of 0.8 for a QTL for a complex trait is shown for pooled DNA designs relative to individual genotyping. The ratio  $N_c/N_{\text{indiv}}$  for affected-unaffected pools (dashed line) is shown as a function the disease incidence  $r$ , while the ratio  $N_{\text{tail}}/N_{\text{indiv}}$  (solid line) is shown as a function of the fraction  $\rho$  of the total population selected for each pool. The optimum value of  $N_{\text{tail}}/N_{\text{indiv}}$  is 1.24, occurring at  $\rho = 27\%$  selected for each pool.

25

Fig. 2a Exact numerical results for the sample size  $N$  required to achieve a type I error rate of  $5 \times 10^{-8}$  with a power of 0.8 are shown for affected-unaffected pools (dashed line) and tail pools (solid line) as a function of the additive variance, or equivalently the genotype relative risk for a heterozygote, for an allele with frequency 0.1 and purely additive variance. Analytic approximations (solid circles), Eqs. 1 and 2, are indistinguishable from the exact results when the genotype relative risk is smaller than a factor of 2. The disease incidence  $r$  is 10% for the affected-unaffected pools, and 27% of the population is selected for the each of the tail pools.

Fig. 2b The frequency difference at the significance threshold is shown for the same parameters as panel *a*. This threshold determines the measurement accuracy required for an association test based on pooled DNA.

### Detailed Description of the Invention

The present invention provides analytic results for association tests. It is shown that the results obtained closely approximate the analytic results to exact numerical calculations. The invention further extends the analysis to qualitative phenotypes using a genotype relative risk model.

A particular quantitative phenotype  $X$  is standardized to have unit variance and zero mean.

The phenotype is hypothesized to be affected by alleles  $A_1$  and  $A_2$ , with frequencies  $p$  and  $1-p$  respectively, at a particular QTL. The population fractions  $P(G)$  for genotypes  $G = A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  are assumed obey Hardy-Weinberg equilibrium. Using standard notation for a variance components model (Falconer and MacKay, 1996), the effect  $\mu_G$  of genotype  $G$  on phenotype  $X$  is  $a-\mu$  for  $A_1A_1$ ,  $d-\mu$  for  $A_1A_2$ , and  $-a-\mu$  for  $A_2A_2$ . The constant  $\mu = (2p-1)a$

+  $2p(1-p)d$  ensures that the mean of  $X$  is zero. The ratio  $d/a$  describes the inheritance mode for allele  $A_1$ . Dominant, recessive, and additive inheritance are special cases with  $d/a$  equal to +1, -1, and 0, respectively.

- 5 The phenotypic variance due to the QTL may be partitioned into the additive variance  $\sigma_A^2$  and the dominance variance  $\sigma_D^2$ , with

$$\sigma_A^2 + \sigma_D^2 = 2pq[a - d(p-q)]^2 + 4p^2q^2d^2.$$

The additive variance is often much larger than the dominance variance even if the inheritance mode is not purely additive. The exceptions are QTLs with a recessive minor alleles and

- 10 dominant major alleles, which are difficult to detect in unselected populations. The contribution of remaining genetic and environmental factors is assumed to follow a normal distribution with residual variance  $\sigma_R^2$ ,

$$\sigma_R^2 = 1 - (\sigma_A^2 + \sigma_D^2).$$

Of particular interest here are complex traits: the effect of any single QTL is small,  $\sigma_A^2 + \sigma_D^2 <$

- 15 0.05, and the residual variance  $\sigma_R^2$  is nearly 1.

A genotype relative risk model corresponds to classifying individuals as affected ( $X > X_T$ ) or

unaffected ( $X < X_T$ ) based on a specific threshold  $X_T$ . The proportion  $r$  of the total population

that is affected is the overall risk or disease incidence; the probability that an individual with

- 20 genotype  $G$  is affected, relative to the probability for an individual with genotype  $A_2A_2$ , is the genotype relative risk. If the inheritance mode of  $A_1$  is additive and  $a$  is small compared to  $\sigma_R$ , the relative risk is multiplicative with allele dose.

The sample size  $N$  required to detect association between genotype  $G$  and the quantitative phenotype or the disease risk depends on the type I error rate  $\alpha$ , the type II error rate  $\beta$ , and the test statistic and experimental design (Snedecor and Cochran, 1989), as well as on the underlying genetic model. For a one-sided test of a single marker,  $\alpha = 1 - \Phi(z_\alpha)$ , where  $\Phi(z)$  is the cumulative probability distribution for standard normal deviate  $z$ , defines  $\alpha$  in terms of deviate  $z_\alpha$ . Similarly,  $1-\beta$  is the power to reject the null hypothesis and  $z_{1-\beta} = \Phi^{-1}(\beta)$ . For a genome scan, the values  $\alpha = 5 \times 10^{-8}$  ( $z_\alpha = 5.33$ ) and  $1-\beta = 0.8$  ( $z_{1-\beta} = -0.84$ ) have been suggested (Risch and Merikangas, 1996).

- 10 We consider two experimental designs using DNA pooled from individuals selected from a sample of size  $N$ : affected-unaffected pools, with DNA pooled from  $n$  affected and  $n$  unaffected individuals; and tail pools, with DNA pooled from  $n$  individuals at each tail of the phenotype distribution. The test statistic for these designs is the frequency difference of the  $A_1$  allele between the pools. The multinomial distribution describing the test statistic may be used
- 15 to calculate exactly the sample size required to achieve statistical significance at specified power.

- When the number of  $A_1$  alleles summed over both pools is large, the distribution of the test statistic is approximately normal. A significant association is detected if the allele frequency
- 20 difference between pools is at least  $z_\alpha$  times the standard deviation of its estimator, or  $z_\alpha p^{1/2}(1-p)^{1/2}/n^{1/2}$ . Furthermore, when the additive variance  $\sigma_A^2$  is small and the residual variance  $\sigma_R^2$  is close to 1, convenient analytic approximations for the sample size requirements may be derived.

For the affected-unaffected design,  $n = rN$  of the individuals are expected to be diagnosed as affected, and an additional  $n$  matched controls are selected from the remainder of the population. The analytic approximation for the sample size is

$$N_{c-c} = [z_\alpha - z_{1-\beta}]^2 [\sigma_R^2 / \sigma_A^2] \cdot 2r(1-r)^2 / y^2 [1 + X_T(1-\sigma_R^2)^{1/2} / 2^{3/2} \sigma_R^2 p^{1/2} (1-p)^{1/2}]^2. \quad (\text{Eq. 1})$$

- 5 The term  $y$  is the height of the standard normal distribution at the normal deviate  $X_T / \sigma_R$  corresponding to the threshold between affected and unaffected phenotypic values.

The tail pools are parameterized by the fraction  $\rho = n/N$  of population selected for each pool, and  $\rho$  plays a role analogous to the overall disease incidence  $r$  in the affected-unaffected

- 10 design. The analytical approximation for the sample size is

$$N_{\text{tail}} = [z_\alpha - z_{1-\beta}]^2 [\sigma_R^2 / \sigma_A^2] \cdot \rho / 2y^2, \quad (\text{Eq. 2})$$

where  $y$  is the height of the standard normal distribution for normal deviate  $\Phi^{-1}(\rho)$ . The design may be optimized by selecting  $\rho$  to minimize  $N_{\text{tail}}$ , which corresponds to minimizing  $\rho / 2y^2$ . With this approximation, the optimal fraction is 0.27 and is independent of  $\alpha$ ,  $\beta$ , and all

- 15 parameters of the genetic model.

A third method, individual genotyping, serves as a baseline for evaluating the efficiency of the two pooling-based methods. The sample size required to achieve significance using individual genotyping is

$$20 \quad N_{\text{indiv}} = [z_\alpha - z_{1-\beta} \sigma_R]^2 / \sigma_A^2, \quad (\text{Eq. 3})$$

based on a regression model of phenotypic value on allele dose.

### Detailed Description of Analytical Methods

The genotype-dependent phenotypic distribution in the variance components model is



$$P(X|G) = (2\pi)^{-1/2} \exp[-(X-\mu_G)^2/\sigma_R^2],$$

and the overall phenotype distribution is the sum of the three normal distributions,

$$P(X) = \sum_G P(X|G)P(G).$$

When an upper threshold  $X_U$  is specified to select a fraction  $\rho$  of the total population with

5 phenotypic values above the threshold, the equation

$$\rho = \sum_G \{1 - \Phi[(X_U - \mu_G)/\sigma_G]\} P(G).$$

may be solved numerically for  $X_U$  as a function of  $\rho$ . The genotypes of individuals selected by

$X > X_U$  follow a multinomial distribution; the probability that an individual has genotype  $G$  is

$$\theta_U(G) = \{1 - \Phi[(X_U - \mu_G)/\sigma_G]\} P(G)/\rho. \text{ A multinomial distribution is similarly defined using a}$$

10 lower threshold  $X_L$ ,

$$1 = \sum_G \theta_L(G) = \rho^{-1} \sum_G \Phi[(X_L - \mu_G)/\sigma_G] P(G).$$

For an affected-unaffected design, the fraction in the upper pool is  $r$  and the fraction in the

lower pool is  $1-r$ , yielding  $X_U = X_L = X_T$ . The relative risk for genotype  $G$  is  $[\theta_U(G)/P(G)]/$

$$[\theta_U(A_2A_2)/P(A_2A_2)].$$

15

Sample size requirements may be obtained directly from the multinomial distributions of

genotypes by exhaustively tabulating allele counts  $C_U$  and  $C_L$  in the upper and lower pools for

each distinct composition of genotypes among the  $n$  selected individuals. The distribution

corresponding to null hypothesis,  $\theta(G) = P(G)$ , is used to define the smallest threshold  $\Delta C$

20 such that  $C_U - C_L \geq \Delta C$  with probability  $\alpha$  or less. The discrete allele count usually yields the

strict inequality. Next, the distributions under the alternative hypothesis are considered, and

the probability that  $C_U - C_L \geq \Delta C$  is tabulated to provide the power. If the power is greater

than or equal to the specified  $1-\beta$ , the choice of  $n$  and  $N = n/\rho$  or  $n/r$  is feasible. A search is



were employed to perform the root search for  $X_U$  and  $X_L$ , the optimization, and the integration over the tail of a normal distribution (Press, 1997).

The analytic results are obtained by setting  $\sigma_1^2$  to  $\sigma_0^2$  and expanding  $\Delta p$  to second order in the effect size  $\mu_G$ , corresponding loosely to a perturbation theory for probability distributions (Chandler, 1987). From a Taylor series expansion,

$$\Phi(z-b) = \Phi(z) - b y - (1/2)b^2 yz,$$

where  $y = (2\pi)^{-1/2} \exp(-z^2/2)$ . Substituting this result into the expressions for  $\mathcal{A}(G)$  using  $b =$

$\mu_G/\sigma_R$  and  $z = X_U/\sigma_R = \Phi^{-1}(1-\rho)$ , where  $X$  is the threshold used to select the pool, yields for the

tail design

$$p_U = p + (y/\rho) E[(\mu_G/\sigma_R)p_G] + (y|z|/2\rho) E[(\mu_G/\sigma_R)^2 p_G] \text{ and}$$

$$p_L = p - (y/\rho) E[(\mu_G/\sigma_R)p_G] + (y|z|/2\rho) E[(\mu_G/\sigma_R)^2 p_G].$$

The corresponding results for the affected-unaffected pools, with  $z = \Phi^{-1}(1-r)$ , are

$$p_U = p + (y/r) E[(\mu_G/\sigma_R)p_G] + (y|z|/2r) E[(\mu_G/\sigma_R)^2 p_G] \text{ and}$$

$$p_L = p - [y/(1-r)] E[(\mu_G/\sigma_R)p_G] - [y|z|/2(1-r)] E[(\mu_G/\sigma_R)^2 p_G].$$

The required expectation values are

$$E[\mu_G p_G] = \Sigma_G P(G) \mu_G p_G = \sigma_A [p(1-p)/2]^{1/2}, \text{ and}$$

$$E[\mu_G^2 p_G] = \Sigma_G P(G) \mu_G^2 p_G = (1/2)(1-\sigma_R^2) - 4p^2(1-p)^2 ad + (2p-1)\sigma_D^2/2 \approx \sigma_A^2/2.$$

The results for  $\Delta p$ ,

$$\Delta p = 2^{1/2} y \sigma_0 \sigma_A / \rho \sigma_R, \text{ tail pools, and}$$

$$\Delta p = [1 + X_T \sigma_A / 2^{3/2} \sigma_0 \sigma_R^2] y \sigma_0 \sigma_A / 2^{1/2} r (1-r) \sigma_R, \text{ affected-unaffected pools,}$$

lead directly to Eqs. 1 and 2.

Approximate genotype relative risks may also be obtained from the Taylor series expansion for  $\theta(G)$ . To lowest order, the relative risk for the heterozygote is approximately  $1 + (d+a)y/r\sigma_R$ , and for the  $A_1A_1$  homozygote is  $1 + 2ay/r\sigma_R$ . For additive inheritance,  $d = 0$ , and the relative risk is multiplicative with allele dose when  $ay/r\sigma_R$  is small. For a complex trait  $\sigma_R$  is close to 1, and for a minor allele,  $a \approx \sigma_A/(2p)^{1/2}$ . When the disease incidence is 10%, the parameter required to be small is  $1.24\sigma_A/p^{1/2}$ .

For individual genotyping, the regression model used to test significance is

$$X = b_1(p_G - p) + \varepsilon,$$

where the residual contribution  $\varepsilon$  to the phenotype has zero mean and is uncorrelated with  $p_G$ .

Using standard statistical methods (Snedecor, 1989), the test statistic  $b_1$  under the null hypothesis has mean zero and variance  $\text{Var}(b_1|\text{null})$  given by

$$\text{Var}(b_1|\text{null}) = N^{-1} \text{Var}(X)/\text{Var}(p_G) = 1/N[p(1-p)/2].$$

Under the alternative hypothesis, the expectation for the test statistic is

$$E(b_1) = \text{Cov}(X, p_G)/\text{Var}(X) = \sigma_A[p(1-p)/2]^{1/2},$$

and its variance is

$$\text{Var}(b_1|\text{alt}) = N^{-1} \text{Var}(\varepsilon)/\text{Var}(p_G) = \sigma_R^2 / N [p(1-p)/2].$$

The sample size required for a one-sided test of  $b_1$  with Type I error  $\alpha$  and power  $1-\beta$  is

$$N = [z_\alpha \text{Var}(b_1|\text{null})^{1/2} - z_{1-\beta} \text{Var}(b_1|\text{alt})^{1/2}]^2 / E(b_1)^2,$$

which is the result provided in Eq. 3.

### Application of the methods of the invention

The sample sizes required for the pooled DNA designs are compared in Fig. 1 to the sample size  $N_{\text{indiv}}$  required by individual genotyping. The ratio  $N_{\text{c-c}}/N_{\text{indiv}}$  (dashed line) is a function of

the disease incidence  $r$ , while  $N_{\text{tail}}/N_{\text{indiv}}$  (solid line) is a function of the pooling fraction  $\rho$ . For typical disease incidence,  $r \sim 10\%$ , the affected-unaffected design requires a sample 5.3× larger than that required for individual genotyping. Compared to the tail design, it measures an allele frequency difference that is half as large and is approximately 4× less efficient. The tail design, with  $\rho = 27\%$ , requires a sample only 1.24× larger than required for individual genotyping. The tail design is also robust to variation in  $\rho$  near its optimum, as values from 19% to 37% drop the efficiency no more than 5%.

The analytic theory indicates that the additive variance  $\sigma_A^2$ , or equivalently the genotype relative risk for an allele of known frequency, is the most important factor determining the sample size requirements. This dependence is shown in Fig. 2a with exact numerical results for affected-unaffected pools (dashed line) and tail pools (solid line) for type I error of  $5 \times 10^{-8}$  and power of 0.8. The minor allele frequency is 10%, its effect on the quantitative phenotype is purely additive, and the disease incidence is 10%. The analytic approximations (solid circles) from Eq. 1 and 2 are nearly indistinguishable from the exact results when the genotype relative risk drops below a factor of 2. As predicted by the analytic theory, the tail pools require smaller sample sizes than the affected-unaffected pools, and the gap grows wider for alleles with a smaller effect on the phenotype. For relative risks of 2 to 5, the deviations from analytic theory are moderate; above a relative risk of 5, the phenotype is monogenic with respect to locus  $G$ , and the analytic approximations for complex traits are no longer valid.

The allele frequency difference between pools at the significance threshold is shown in Fig. 2b for affected-unaffected pools (dashed line) and tail pools (solid line). The measurement error in the allele frequency difference must be smaller than the significance threshold to detect

association (Darvasi, 1994). Evaluations that provide a frequency difference measurement accurate to 0.04 can detect association with alleles responsible for 1% of the total phenotypic variance, corresponding to a heterozygote relative risk of 1.5. The allele frequency difference measurement must be accurate to 0.01 to detect association with an allele explaining 0.1% of the phenotypic variance, corresponding to a relative risk of 1.14.

To test the range of validity of the analytic estimates for pooling, we performed a series of exact calculations of sample size requirements as a function of  $p$  and  $d/a$ . Large deviations were seen only when the magnitude of a gene effect  $\mu_G$  approached  $\sigma_R$  in size, or, equivalently, when  $\sigma_A^2$  was larger than the minor allele frequency or when a genotype relative risk was larger than 5 (results not shown). For additive contributions from a minor allele, the range of validity corresponds to  $\sigma_A^2 < 2p$ .

The advantages of the methods disclosed herein include the following. The optimal fraction for tail pooling, 27%, is independent of all model parameters including allele frequency, inheritance mode, effect size, and type I error and power, for virtually any QTL contributing to a complex trait. The exceptions to this finding are rare QTLs with relative risks of 5 or greater, and rare, recessive alleles, both of which are more difficult to detect than more frequent alleles contributing to the same overall phenotypic variance. In addition, the tail design is approximately 4-fold more efficient than the affected-unaffected design and requires a sample size only 24% larger than for individual genotyping. Still further, DNA pooling studies designed according to the present procedures disclosed herein provide extremely efficient methods for large-scale screening and should help to make feasible genome-wide association studies.

## References

- Abecasis, GR, Cardon, LR, Cookson, WOC (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66: 279-292.
- 5 Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999 Jul;22(3):231-238.
- 10 Collins A, Lonjou C, Morton NE (2000) Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* 96:15173-15177.
- Daniels, J.K., Holmans, P., Williams, N.M., Turic, D., McGuffin, P., Plomin, R., Owen, M.J. A simple method for analysing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am. J. Hum. Genet.* 62, 1189-1197 (1998).
- 15 Darvasi A, Soller M (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138: 1365-1373.
- 20 Falconer, D.S., and MacKay, T.F.C. *Introduction to quantitative genetics*. (Addison-Wesley, Boston, 1996).
- Frank, L (2000) Storm brews over gene bank of Estonian population. *Science* 286: 1262.
- 25 Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* 56:1224-1233.
- Fulker, D.W., Cherny, S.S., Sham, P.C., Hewitt, J.K. Combined linkage and association analysis of quantitative traits. *Am. J. Hum. Genet.* 64, 259-267 (1999).
- 30 Hill, W. G. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* 27, 293-311 (1971).
- Kimura, M. & Crow, J.F. Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl. Acad. Sci. USA* 75, 6168-6171 (1978).
- 35 Kruglyak, L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22: 139-144.
- 40 Liu, B-H (1997) *Statistical Genomics*. CRC Press, Boca Raton.
- Nilsson A, Rose J (1999) Sweden takes steps to protect tissue banks. *Science* 286: 894.
- Ott J (1999) *Analysis of human genetic linkage*. Johns Hopkins Univ Pr, Baltimore.
- 45 Rabinow, P (1999) *French DNA: Trouble in Purgatory*. University of Chicago Press, Chicago.

Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847-856 (2000).

- 5 Risch NJ, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516-1517.

- Risch NJ, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*  
10 8:1273-1288.

Sham, P (1997) *Statistics in Human Genetics*. Arnold.

- 15 Sham, P.C., Cherny, S.S., Purcell, S., Hewitt, J.K. Power of linkage versus association analysis of quantitative traits, by use of variance components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616-1630 (2000).

- 20 Snedecor, G. W., and Cochran, W. G. *Statistical Methods, Eighth Edition*. (Iowa State University Press, Ames, 1989).

Beyer, W.H. (ed). *CRC Standard Mathematical Tables, 27<sup>th</sup> Edition*. (CRC Press, Boca Raton, FL, 1984).

- 25 Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. *Numerical Recipes in C, The Art of Scientific Computing, Second Edition* (Cambridge University Press, Cambridge, UK, 1997).

- Chandler, D. *Introduction to Modern Statistical Mechanics*. (Oxford Univ. Press, New York, 1987).

- 30 Ollivier, L., Messer, L.A., Rothschild, M.F. & Legault, C. The use of selection experiments for detecting quantitative trait loci. *Genet. Res., Camb.* **69**, 227-232 (1997).



### OTHER EMBODIMENTS

While the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the  
5 invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.